Understanding User Satisfaction with Intelligent Assistants

Julia Kiseleva[1,*] Kyle Williams[2,*] Jiepu Jiang[3,*] Ahmed Hassan Awadallah[4]
Aidan C. Crook[4] Imed Zitouni[4] Tasos Anastasakos[4]

[1]Eindhoven University of Technology, j.kiseleva@tue.nl
[2]Pennsylvania State University, kwilliams@psu.edu
[3]University of Massachusetts Amherst, jpjiang@cs.umass.edu
[4]Microsoft, {hassanam, aidan.crook, izitouni, tasos.anastasakos}@microsoft.com

ABSTRACT

Voice-controlled intelligent personal assistants, such as Cortana, Google Now, Siri and Alexa, are increasingly becoming a part of users' daily lives, especially on mobile devices. They introduce a significant change in information access, not only by introducing voice control and touch gestures but also by enabling dialogues where the context is preserved. This raises the need for evaluation of their effectiveness in assisting users with their tasks. However, in order to understand which type of user interactions reflect different degrees of user satisfaction we need explicit judgements. In this paper, we describe a user study that was designed to measure user satisfaction over a range of typical scenarios of use: controlling a device, web search, and structured search dialogue. Using this data, we study how user satisfaction varied with different usage scenarios and what signals can be used for modeling satisfaction in the different scenarios. We find that the notion of satisfaction varies across different scenarios, and show that, in some scenarios (e.g. making a phone call), task completion is very important while for others (e.g. planning a night out), the amount of effort spent is key. We also study how the nature and complexity of the task at hand affects user satisfaction, and find that preserving the conversation context is essential and that overall task-level satisfaction cannot be reduced to query-level satisfaction alone. Finally, we shed light on the relative effectiveness and usefulness of voice-controlled intelligent agents, explaining their increasing popularity and uptake relative to the traditional query-response interaction.

Keywords: intelligent assistant, user satisfaction, user study, user experience, mobile search, spoken dialogue system

1. INTRODUCTION

Spoken dialogue systems [35] have been around for a while. However, it has only been in recent years that voice controlled intelligent assistants, such as Microsoft's Cortana, Google Now, Apple's Siri, Amazon's Alexa, Facebook's M, etc, have become a daily used feature on mobile devices. A recent study [12], executed by Northstar Research and commissioned by Google, found out that 55% of the U.S. teens use voice search every day and that 89% of teens and 85% of adults agree that voice search is going to be 'very common' in the future. One of the reasons for the increased adoption is the current quality of speech recognition due to massive online processing [36], but perhaps more important is the added value users perceive: the spoken dialogue mode of interaction is a more natural way for people to communicate and is often faster than typing.

CHIIR '16, March 13 - 17, 2016, Carrboro, NC, USA

ISBN 978-1-4503-3751-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2854946.2854961

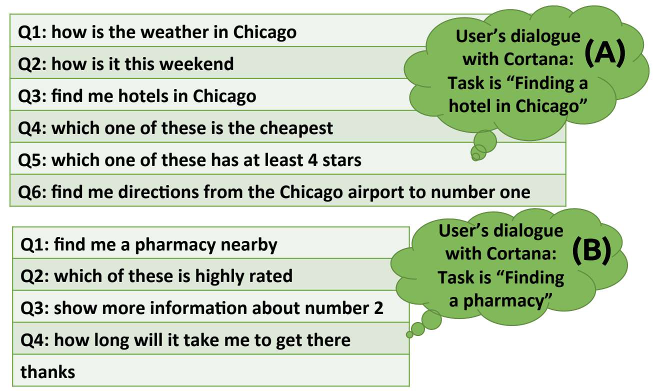

Figure 1: Two real examples of users' dialogues with a intelligent assistant: In the dialogue (A), a user performs a 'complex' task of planning his weekend in Chicago. In the dialogue (B), a user searches for the closest pharmacy.

Intelligent assistants enable new mechanisms of information access, that are very different from traditional web search. Figure 1 shows two examples of dialogues with intelligent assistants sampled from the interaction logs. They are related to two tasks: **(A):** searching things to do on a weekend in Chicago, and **(B):** searching for the closest pharmacy. Users express their information needs in spoken form to an intelligent assistant. The user behavior is different compared with standard web search because in this scenario an intelligent assistant is expected to maintain the context throughout the conversation. For instance, our user anticipates intelligent assistants to understand that their interaction is about 'Chicago' in the transitions: $Q_1 \rightarrow Q_2$, $Q_3 \rightarrow Q_4$ in Figure 1(A). These structured search dialogues are more complicated than standard web search, resembling complex, context-rich, task-based search [43]. Users expect their intelligent assistants to understand their intent and to keep the context of the dialogue—some users even *thank* their intelligent assistant for its service, as in example in Figure 1(B).

Users communicate with intelligent assistants through voice commands for different scenarios of use, ranging from controlling their device—for example to make a phone call, or to manage their cal-

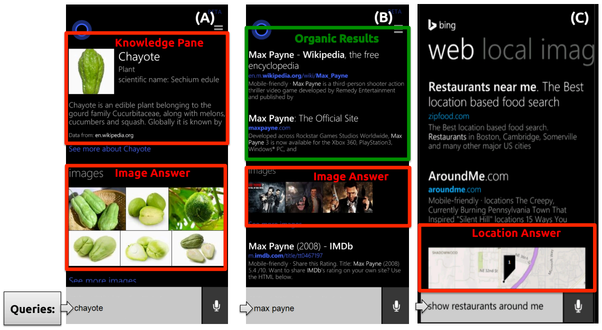

Figure 2: An example of mobile SERPs that might lead to 'good abandonment'.

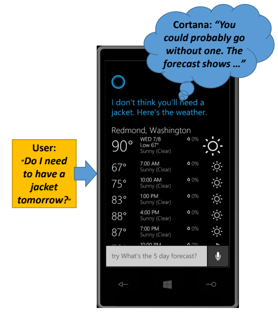

Figure 3: An example of a 'simple' task with a structured search dialogue.

endar—to complex dialogues as shown in Figure 1. These interactions between users and intelligent assistants are more complicated than web search because they involve:

- automatic speech recognition (ASR): users communicate mostly through voice commands and it has been shown that errors in speech recognition negatively influence user satisfaction [23];
- understanding user intent: an intelligent assistant needs to understand user intent in order to take action on the intended task, or to provide an exact answer when possible;
- dialogue-based interaction: users expect an intelligent assistant to maintain the context of the dialogue;
- complex information needs: users express more sophisticated information needs while interacting with intelligent assistants.

This prompts the need to better understand success and failure of intelligent assistant usage. When are users (dis)satisfied? How can we evaluate intelligent assistants in ways that reflect perceived user satisfaction well? Can we resort to traditional methods of offline and online evaluation or do we need to take other factors into consideration?

Evaluation is a central component of many web search applications because it helps to understand which direction to take in order to improve a system. The common practice is to create a 'gold' standard (set of 'correct' answers) judged by editorial judges [21]. In case of intelligent assistants, there may be no general 'correct' answer since the answers are highly personalized and contextualized (e.g., by the user's location, prior queries or interactions) to fit user information needs. Another way to evaluate web search performance is through implicit relevance feedback such as clicks and dwell time [3, 11, 16, 26, 27]. However, we know that user satisfaction for mobile web search is already very different [33].

In the examples in Figure 2, different types of answers are shown for queries such as *'Location Answer'*, *'Image Answer'* or *'Knowledge Pane Answer'*. Users can find required information directly on the search result page (SERP) and they do not need to perform any further interactions (e.g. clicks). So we cannot assume that users who do not interact with the SERP are dissatisfied. This problem of 'good' abandonment received a lot of interest in recent years [6–8, 34]. An example of a users' dialogue about *'weather'* is shown in Figure 3. All information about the weather is already shown to the users and they do not need to click. In case of structured dialogue search, the lack of standard implicit feedback signals emerges even more because users talk to their phones instead of making clicks. One example of this is the transition $Q_2 \rightarrow Q_3$ in Figure 1(B).

In light of the current work, this paper aims to answer the following main research question:

> *What determines user satisfaction with intelligent assistants?*

We breakdown our general research problem into five specific research questions. Our first research question is:

> **RQ 1:** *What are characteristic types of scenarios of use?*

Based on analysis of the logs of a commercial intelligent assistant; and from previous work [25], we propose three types of scenarios of intelligent assistant use: **(1)** controlling the device; **(2)** searching the web; and **(3)** perform a complex (or 'mission') task in a dialogue interaction. We characterize key aspects of user satisfaction for each of these scenarios.

Our second research question is:

> **RQ 2:** *How can we measure different aspects of user satisfaction?*

We set up user studies with realistic tasks derived from the log analysis, following the three scenarios of use, and measuring a wide range of aspects of user satisfaction relevant to each specific scenario.

Our third research question is:

> **RQ 3:** *What are key factors determining user satisfaction for the different scenarios?*

In order to understand what the key components of user satisfaction are, we analyze output of our user studies for different intelligent assistants scenarios. We aim at understanding what factors influence user satisfaction the most: speech recognition quality, complexity of the task, or the amount of effort required to complete the task.

Our fourth research question is:

> **RQ 4:** *How to characterize 'abandonment' in the web search scenario?*

'Good abandonment' makes it difficult to measure user satisfaction with web search scenario using conventional implicit feedback behavioral signals. We analyze the way in which users interact with the intelligent assistant following a web search; we characterize user satisfaction in general, and over the number of issued queries, and types of answers found.

Our fifth research question is:

RQ 5: *How does query-level satisfaction relate to overall user satisfaction for the structured search dialogue scenario?*

The structured search dialogue scenario introduced a new mechanism for users to interact with intelligent assistants which has not received a lot of attention in the literature. We analyze the data for the search dialogue interactions, and investigate satisfaction over tasks with increasing complexity; we consider how sub-task level satisfaction relates to overall task satisfaction.

The remainder of this paper is organized as follows. Section 2 describes earlier work and background. Then, Section 3 introduces scenarios of user interaction with intelligent assistants, discusses differences and similarities in user behavior. Section 4 describes different types of user studies developed to evaluate user satisfaction for intelligent assistants different scenarios. Finally, Section 5 reports our results and findings. We summarize our findings, discuss possible extensions of the current work in Section 6.

2. RELATED WORK

In this section, we will discuss related work relevant to the research described in this paper, covering three broad strands of research. First, methods for evaluating user satisfaction in web search systems are presented in Section 2.1. Research on spoken dialogue systems is discussed in Section 2.2. Finally, we focus on user studies for the evaluation of intelligent assistants in Section 2.3.

2.1 Evaluating User Satisfaction

User behavioural signals have been extensively studied and used for the evaluation of web search systems [1, 2, 14–16, 24, 30, 45]. Historically, the key objective of information retrieval systems is to retrieve relevant information (typically documents) or references to documents containing required information [37, 38]. Given this query-document relevance score, many metrics have been defined: MAP, NDCG, DCG, MRR, P@n, TBG, etc. [21]. For such setup we have a collection of documents and queries that are annotated by human judges. It is a common setup used at TREC[1]. In this case we evaluate system performance at the *query-level* for the pair $\langle Q, SERP \rangle$. Building such data collections needed for this type of evaluation is both expensive and time consuming. There is a risk that such collections may be noisy, given that third-party annotators have limited knowledge of an individual user intent.

User satisfaction is widely adopted as a subjective measure of search experience. Kelly [28] proposes a definition: *'satisfaction can be understood as the fulfillment of a specified desire or goal'*. Furthermore, recently researchers studied different metrics reflective of user satisfaction such as effort [48] and it has been shown that user satisfaction at the query-level can change over time [31, 32] due to some external influences. These changes lead to the necessity of updating the data collection. Unfortunately, *query-level* satisfaction metrics ignore the information about a user's 'journey' from a question to an answer which might take more than one query [22]. Al-Maskari et al. [4] claim that *query-level* satisfaction is not applicable for informational queries – users can run follow-up queries if they are unsatisfied with the returned results; reformulations can lead users to an answer; this scenario is called *task-level* user satisfaction [9, 16]. Previous research proposed different methods for identifying successful sessions: Hassan et al. [16] used a Markov model to predict success at the end of the task; Ageev et al. [1] exploited an expertise-dependent difference in search behavior by using a Conditional Random Fields model to predict a search success – authors used a game-like strategy for collecting annotated data by asking participants to find answers to non-trivial questions using web search. On the other hand, situations when users are frustrated have also been studied: Feild et al. [10] proposed a method for understanding user frustration. Hassan et al. [17] and Hassan Awadallah et al. [18] have found that high similarity of queries is an indicator of an unsuccessful task. All described methods focus on analyzing user behavior when users interact with traditional search systems.

[1]Text REtrieval Conference: http://trec.nist.gov/

2.2 Spoken Dialogue Systems

The main difference between traditional web search and intelligent assistants is their conversational nature of interaction with users. In the considered scenarios of usage of intelligent assistants, the technology can refer to the users' previous requests in order to understand the context of a conversation. For instance, in the dialogue (A) in Figure 1, the user asks for Q_2 and assumes that the intelligent assistant will 'remember' that he is interested in Chicago. Therefore, the spoken dialogue systems [35] are closely related to intelligent assistants because the spoken dialogue systems understand and respond to the voice commands in a dialogue form. This area has been studied extensively over the past two decades [40–42]. Most of these studies focused on systems that have not been deployed in a large scale and hence did not have the necessary means to study how users interact with these systems in real-world scenarios. However, intelligent assistants are different from traditional spoken dialogue systems because they also support interactions and 'understand' user intent. Furthermore, intelligent assistants display an answer which users can interact with and they are not purely based on speech—users can type in responses as well. From these perspectives, intelligent assistants are similar to multi-modal conversational systems [19, 44].

2.3 User Studies of Intelligent Assistants

In recent years voice-controlled personal assistants have become available to the general public. There are few studies researching intelligent assistants, and there is only one earlier paper that organizes a user study [25]. Jiang et al. [25] focus on simulated tasks for device control, as well as chat and web search, and identify satisfactory and unsatisfactory sessions based on features used in predicting satisfaction on the web, as well as acoustic features of the spoken request. Our work extends this study focusing on a wider range of scenarios of intelligent assistant use, including complex dialogues, and analyzing crucial aspects determining user satisfaction under these different conditions.

More broadly, intelligent assistants are often used for longer sessions and tasks that involve sub-tasks and complex interactions, and task complexity has been studied in many user studies. Wildemuth et al. [46] reviewed over a hundred interactive information retrieval studies in terms of task complexity and difficulty, and found that the number of sub-tasks, the number of facets, and the indeterminably were the main dimensions of task complexity. The structured search tasks we use in our study score high on these dimensions. Recently, Kelly [29] linked perceived task complexity with effort, suggesting that user satisfaction may depend on the amount of effort required to complete a complex task. We also look specifically at the role of effort relative to task-level user satisfaction.

To summarize, the key distinctions of our work compared to previous efforts are: we studied how users interact with intelligent assistants; we studied how we can use these interactions to understand *'good abandonment'*; we explored three main scenarios of user interactions with intelligent assistants and a definition of user satisfaction for these scenarios.

3. USER INTERACTION WITH INTELLIGENT ASSISTANTS

This section reports our study findings pertaining to the **RQ 1:** *What are characteristic types of scenarios of use?* In order to answer our research question we used the Microsoft intelligent assistant — Cortana. Historically, the scenario of controlling devices through voice commands was implemented first. It is described in detail in Section 3.1. From a user-satisfaction perspective, the main difference of this scenario compared with an information seeking task is that the 'right answer' is clear; in order to satisfy a user, an intelligent assistant needs to interpret requests correctly and give access to the correct functionality. In contrast, for information seeking tasks [20, 47] users exhibit different behaviour. Cortana responds to a general search scenario by returning a variant of the Bing Mobile SERP, which may include answers or tiles from the knowledge pane as well as organic search results (see Figure 2); we discuss this scenario in Section 3.2. Another mechanism by which users interact with information systems that some intelligent assistants support is the 'structured search dialogue' (Figure 1). In this case, intelligent assistants are able to maintain the context of a conversation as the system engages with the user in a dialogue; it is definitely more complex (for the system) but at the same time more natural (for the user) form of 'communication' between users and information systems. This scenario is presented in Section 3.3.

3.1 Controlling a Device

The first scenario of using intelligent assistants that we study is the direct access of on-device functionality – e.g., call a contact, check the calendar, access an app, etc. This scenario is useful because, ordinarily, it takes several actions to complete on existing smartphones. For example, in order to make a phone call, the user needs to first access a contact list on the phone and then identify the desired person. The ordinary process is time consuming, especially when the user is not familiar with the device. Instead, one can directly talk to the intelligent assistant to solve the problem, e.g., 'call Sam'. As long as the intelligent assistant can correctly recognize the user's words and task context, this largely reduces the user's effort.

Our user study includes the following types of on-device tasks that are popular in Cortana's usage logs:

- Call a person;
- Send a text message;
- Check on-device calendar;
- Open an application;
- Turn on/off wi-fi;
- Play music.

We group these tasks into one category because they share the similarity that users try to access these on-device functions through the intelligent assistants. These functions are normally not provided by the intelligent assistants, but offered by the device hosting it. In these tasks, intelligent assistants serve as a quick and efficient interface for accessing on-device functionality.

3.2 Performing Mobile Web Search

Another popular usage scenario for intelligent assistants is the general web search scenario. For this scenario, input can be either speech or text and there is no need for the system to be state-aware since it does not provide a multi-turn experience. During web search on mobile devices, the intent can be ambiguous. Therefore, the search result page (SERP) is very diverse and may include different types of answers such as:

- *'Answer Box'*. A box such as the knowledge pane (Figure 2(A)) or directions to a location (Figure 2(C)). These answer boxes are present for specific query intents.
- *'Image'*. In this case, just seeing an image may have satisfied a user's information need (e.g. Figure 2(A,B)).
- *'Snippet'*. The user's information need is satisfied by a snippet of text appearing below an organic search result (e.g. Figure 2(B)).

These different elements on a SERP can all lead to user satisfaction. For instance, the knowledge pane might contain the answer that the user is looking for or a user may be satisfied by the text in a snippet.

In some cases, the SERP is able to directly satisfy the user's information need and it can lead to the absence of one of the most studied user interaction signals (i.e. clicks on the SERP). Previous work on general web search has shown that presenting these types of answers affects user behavior [33] and leads to 'good abandonment' [7, 34] where the user appears to have abandoned the results but was actually satisfied without the need to engage with the SERP using clicks.

3.3 Structured Search Dialogue

In the structured search dialogue scenarios, the users are engaged in a conversation with the system using voice as we show in Figure 1. Cortana returns a structured answer that is distinguishably different from the usual SERP (Figure 2). The key component of this scenario is the ability of the intelligent assistant to maintain the context of the conversation. Examples of tasks where this scenario is activated include places (e.g. restaurants, hotels, travel, etc.) and weather. There are two types of tasks that fall under this scenario: *'simple'* and *'mission'* tasks. We discuss *'simple'* tasks in Section 3.3.1 and *'mission'* tasks in Section 3.3.2.

3.3.1 Simple Tasks

'Simple' tasks have one underlying atomic information need and mostly consist of one query and one answer. An example of a 'simple' task is the weather-related task shown in Figure 3. 'Simple' tasks can be very similar to web search scenarios. We expect that they can be evaluated using a paradigm of *query-level* satisfaction because 'simple' task usually consists of one query and one answer.

3.3.2 Mission Tasks

'Mission' tasks consist of multiple interactions with Cortana that lead towards one final goal (e.g. 'find a place for vacation'). The final task can be divided into sub-tasks; the complexity of 'missions' is dependent on the need to understand the context of the conversation.

The example of a places-related 'mission' dialogue is presented in Figure 4. A user makes the following transitions:

- **(1)** 'asking for a list of the nearest restaurant' $\rightarrow$ **(2)** 'sorting the derived list to find best restaurants';

  *(**Comment for the transition** $1 \rightarrow 2$ **:** Cortana 'knows' that a user is working on the same list of restaurants)*

- **(2)** $\rightarrow$ **(3)** 'selecting the restaurant from the list and asking for the directions';

  *(**Comment for the transition** $2 \rightarrow 3$ **:** Cortana 'knows' that a user is working with the sorted list of restaurants)*

Figure 4: An example of a structured search dialogue (mission task).

This type of interaction can be viewed as a sequence of user requests ('user journey towards a information goal') where each request is a step towards user satisfaction or frustration. Much of the frustration happens when Cortana is not able to keep the context and users need to re-attempt the task from the start. Going back to the example in Figure 1 (B), if Cortana did not carry the context across the transition $Q_3 \rightarrow Q_4$ (e.g. due to ASR error) then the user has to restart the task. Overall, user satisfaction goes down dramatically in this case, especially because the mistake happens at the end of the session.

To summarize, in this section, we categorized three distinct scenarios of user interactions with intelligent assistants. Cortana was used as an intelligent assistant example. We discussed difficulties in evaluating user satisfaction in each of these scenarios. For the controlling a device scenario, users' requests cannot be characterized by information needs. In order to satisfy users' needs the system is required to recognize their speech correctly and map a request to the right functionality. The web search and structured search dialogue are more complex because a comprehensive information seeking process is involved. The effect of good abandonment makes it difficult to measure user satisfaction. The structured search dialogue is a novel way of users' interactions that support complex tasks which consist of more than one singular objectives. We refer to these complex tasks as 'mission' tasks.

4. DESIGNING USER STUDIES

This section addresses **RQ 2:** *How can we measure different aspects of user satisfaction?* by describing the design of user study to collect user interaction data and ratings for different intelligent assistant scenarios. We start by characterizing the participants of our study in Section 4.1 followed by a description of the environment of the studies in Section 4.2. The general procedure for the study is presented in Section 4.3. Then, we present the detailed tasks and user study procedure for the different scenarios separately: device control in Section 4.4, structured search dialogue in Section 4.6, and mobile web search in Section 4.5. While designing the user study tasks we follow two requirements: **(1)** the simulated tasks should be realistic and as close as possible to real-world tasks ; **(2)** according to Borlund [5] we construct the simulated tasks so that participants could relate to them and they would provide '*enough imaginative context*.'

4.1 Participants

We recruited 60 participants through emails sent to a mailing list of an IT company located in the United States. All participants were college or graduate students interning at the company or full time employees. They are reimbursed $10 gift card for participating in an experiment. The average age of participants is 25.53 (± 5.42). The characteristics of participants regarding gender (A), field of education (B) and native language (C) are presented in Table 1.

Table 1: Demographics of the user study participants: gender (A), native language (B), and field of education (C)

Gender		Native language		Field of education	
Male	75%	*English*	55%	*Computer science*	82%
Female	25%	*Other*	45%	*Electrical engineering*	8%
				Mathematics	7%
				Other	2%

4.2 Environment

Participants performed the tasks on a Windows phone with the latest version of Windows Phone 8.1 and Cortana installed. If the task needed to access some device resources, functions or applications (e.g. maps), they are installed to make sure users would not encounter problems. The experiment was conducted in a quiet room, so as to reduce the disturbance of environment noise. Although the real environment often involves noise and interruption, we eliminate those factors to simplify the experiment.

4.3 General Procedure

The participants were first asked to watch a video introducing the different usage scenarios of Cortana, and then complete a background questionnaire with demographics and previous experience with using intelligent assistants. Then, they work on one training task and eight formal tasks. We instructed participants that they could stop a task when they had accomplished the goal or if they became frustrated and wanted to give up. Finally, they were asked to answer an extensive questionnaire on their experience and share further details during a short interview.

For each task, we asked participants to listen to an audio recording that verbally described the task objective. We did not show the participants the task description while they were working on the task, because in an earlier pilot study, many participants directly used the sentences shown in task descriptions as requests. We strongly want to avoid such outcome because our goal is to simulate real user behavior. After completing the task, participants were directed to the questionnaires. The questions depend on the objectives of the experiment and vary per user study. Participants answered all questions using a standard 5-point Likert scale.

4.4 User Study for Controlling Device

The first user study is to conduct the most basic scenario–controlling a device. We will now describe the tasks (Section 4.4.1) and the specific procedure for this study (Section 4.4.2).

4.4.1 Tasks

In total we develop nine device control tasks. We rotated the assignment of tasks using a Latin square such that 20 participants worked on each unique task. Some examples of these tasks are:

- *Ask Cortana to play a song by Michael Jackson (a song by the artist is downloaded on the device prior to the task).*
- *You are on your way to a meeting with James, but will be late due to heavy traffic. Send James Smith a text message using Cortana and explain your situation.*

- *Create a reminder for a meeting with James next Thursday at 3pm.*
- *Ask Cortana to turn off the Wi-Fi on your phone.*
- *Ask Cortana to open WhatsApp (the name of a popular App, and the App is installed on the device prior to the task).*

4.4.2 Procedure

The instructional video about the controlling device scenario is about 2 minutes long. Our informal observation is that the video instructions were effective and felt like a natural extension of the speech interaction of the study, framing the study for the participants better than written instruction would do. When the participants worked on this user study, they were asked to use mostly voice for interactions. After terminating a task, they answer questions regarding their experience, including:

1. *Were you able to complete the task?*
2. *How satisfied are you with your experience in this task?*
3. *How well did Cortana recognize what you said?*
4. *Did you put in a lot of effort to complete the task?*

The total experiment time was about 20 minutes.

4.5 User Study for Web Search

The next use-case for the user study is general web search. There has already been significant research involving search on mobile phones [33, 39]; however, *'good abandonment'* in mobile search has had limited investigation. It is a particularly interesting problem to investigate as queries in mobile search have been described as *quick answer types* and previous research has shown that users formulate mobile queries in such as way so as to increase the likelihood of the query being satisfied directly on the SERP [34]. For this reason, in this user study we choose to focus on tasks that have an increasing likelihood of leading to good abandonment. Section 4.5.1 introduces the used tasks . The specific procedure for this study is presented in Section 4.5.2.

4.5.1 Tasks

The tasks for web search were designed to encourage answer-seeking behavior and increase the likelihood of good abandonment. The tasks involved:

- *A conversion from the imperial system to the metric system.*
- *Determining if it was a good time to phone a friend in another part of the world.*
- *Finding the score of the user's favourite sports team.*
- *Finding the user's favourite celebrity's hair colour.*
- *Finding the CEO of a company that lost most of its value within the last 10 years.*

After data cleaning, we retained the data from 55 users who completed a total of 274 tasks, 194 of which were labeled as SAT, while the remaining 70 were labeled as DSAT. There were a total of 607 queries for these tasks of which 576 were abandoned, thereby indicating that we were successful in designing tasks that had a higher potential of leading to good abandonment.

4.5.2 Procedure

The user study starts with the instructional video (about 3 minutes long) that contains an example task for general web search. After completing each task, users were asked:

1. *Were you able to complete the task?*
2. *Where did you find the answer?*

   (**Suggested Answers:** *In an answer box*; *On a website that I visited*; *In a search result snippet*; *In an image*.)
3. *Which query led you to finding the answer?*

   (**Suggested Answers:** First; Second; Third; Fourth or later)
4. *How satisfied are you with your experience in this task?*
5. *Did you put in a lot of effort to complete the task?*

The purpose of the second question was to allow us to better understand where users find information that they are looking for. The option *'On a Website that I visited'* means a user clicked on a search result and visited a website to find the information that they were looking for.

The purpose of the third question was to allow us to tie a success event within a task to a specific query for future evaluation. We did not ask users about ASR quality because we gave users the option of using text input instead of speech. The reason for doing this is that, since we wanted to study good abandonment, we tried to reduce the level of frustration due to speech recognition errors. However, even though that was the case, we still found that most of the participants used voice input because they found it more convenient. The total experiment time was about 20 minutes.

4.6 User Study for Structured Search Dialogue

This Section introduces the design of the user study to explore user satisfaction for the structured search dialogue. First, we describe the way we create tasks for our user study and tasks examples in Section 4.6.1. The specific procedure for this study is described in Section 4.6.2.

4.6.1 Tasks

In order to come with the list of tasks for participants, Cortana's logs (over 400K requests) are analyzed. We look at the terms distribution to get an idea for what kind of places users are looking for. Based on our analysis we come up with eight tasks that designed to cover a large portion of topics used by Cortana's users.

Among these eight tasks we have:

- *(A) one simple task that is related to the weather where almost all participants are satisfied;*
- *(B) four 'mission' tasks that include two sub-tasks;*
- *(C) three 'mission' tasks that require at least three switches in a subject.*

Tasks are given to participants in a free/general form in order to get query diversity and stimulate use satisfaction or frustration with returned results. For instance, let us consider the 'mission' task with 3 sub-tasks: *'You are planning a vacation. Pick a place. Check if the weather is good enough for the period you are planning the vacation. Find a hotel that suits you. Find the driving directions to this place.* By giving a free-form task we stimulate the information need of participants (they need to come up with their own goal and they are more involved in the tasks) so this scenario should lead to satisfaction or frustration. For instance, out of 60 responses for the described task we get 46 unique places.

As a result of free-form task-formulation we obtained a diverse query set, characterized by the following: participants performed a total of 540 tasks that incorporated 2,040 queries, of which 1,969 were unique and the average query-length is 7.07. The simple task generated 130 queries in total; five (B)-type tasks generated 685 queries; three (C)-type tasks generated 1,355 queries.

4.6.2 Procedure

The introductional video for this user study is about 4 minutes long and informs participants how to use the structured search dialogue. During this user study, we instruct participants to verbally interact with Cortana. We instruct them to use text input only if Cortana does not understand their requests more than three times. Only after completing a task are they then redirected to questions regarding their experience in this task session. For 'mission' tasks, users are asked to indicate their satisfaction with both the sub-tasks and the whole task in general. In order to stimulate participant involvement in the tasks, we asked them to answer clarifying questions. For instance, if the task was *'what is the weather tomorrow'*, the user also needed to indicate the temperature; this way we keep participants engaged.

Participants answer the following four questions after completing the tasks:

1. *Were you able to complete the task?*
2. *How satisfied are you with your experience in this task in general?*

   If the task has sub-tasks participants indicate their graded satisfaction e.g. **a.** *How satisfied are you with your experience in finding a hotel?* **b.** *How satisfied are you with your experience in finding directions?*
3. *Did you put in a lot of effort to complete the task?*
4. *How well did Cortana recognize what you said?*

The total experiment time was about 30 minutes.

To summarize, we described how we designed user study with the objective of understanding user satisfaction with different scenarios of intelligent assistants, measuring relevant variables as speech recognition quality, task completion, and the effort taken. The introductory videos designed for the user study are available.[2] Detailed descriptions of the tasks and the recording on the task can be accessed.[3]

5. RESULTS AND FINDINGS

This section presents the results and findings from the user studies, investigating our three remaining research questions (RQ3–5). In Section 5.1, we focus on the user satisfaction relative to the different usage scenarios, and in relation to other measures like the speech recognition, task completion and effort taken. In Section 5.2, we analyze 'good abandonment' in web search, in short sessions where answers may be shown without the need for further interaction. In Section 5.3, we focus on structured search dialogues and how session- or task-level satisfaction relates to subtask-level satisfaction for longer sessions.

5.1 Scenarios of Use

[2]https://goo.gl/6Gv5Y5

[3]https://goo.gl/0jXu2J

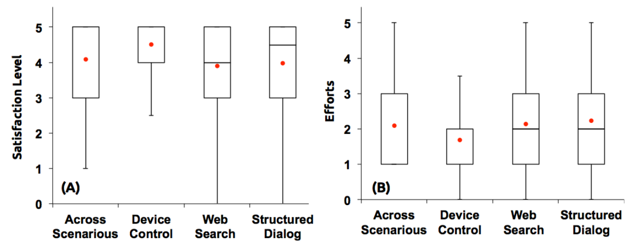

Figure 5: User satisfaction (A) and effort (B) across scenarios and in three discussed scenarios separately. Mean is red dot. Median is horizontal line.

Table 2: Correlations of user satisfaction with other measures: ASR quality, Task Completeness, User Efforts. The sign * stands for statistically significant results ($p < 0.05$)

Measures	All	Device Control	Web Search	Struct. Dialogue
SAT vs. ASR	0.57^*	0.57	$-^\dagger$	0.56^*
SAT vs. Completion	0.18^*	0.59^*	0.10	0.10^*
SAT vs. Effort	-0.75^*	-0.64^*	-0.65^*	-0.80^*
ASR vs. Completion	-0.22^*	-0.27^*	$-^\dagger$	-0.19^*
ASR vs. Effort	-0.54^*	-0.56^*	$-^\dagger$	-0.51^*
Completion vs. Effort	-0.11^*	-0.39^*	-0.08^*	-0.05^*

†ASR was not calculated for web search as both spoken and typed queries were used.

We will now investigate **RQ 3:** *What are key factors determining user satisfaction for the different scenarios?* The scenarios of use differ considerably in terms of complexity, session duration, type of outcome, and more, suggesting that different factors may play a role in determining user satisfaction.

We first discuss the distribution of user satisfaction across all aforementioned mechanisms of intelligent assistant use, both over the entire session and broken down by scenario—device control, web search on a mobile device, and structured search dialogue — which is presented in Figure 5(A). The user satisfaction is very high with means around 4 on a 5-point scale, both overall sessions and for each of the three scenarios. The high level of satisfaction showcases the maturity of the current generation of intelligent assistants, and explains the increasing adoption. As a case in point, many participants had (almost) never used the service, and were impressed by its effectiveness. We can see that user satisfaction with the device controlling tasks (mean of 4.5) is somewhat higher on average than with the information seeking tasks (mean of 3.7), plausibly because the information seeking tasks are open domain and more complex.

We also show the distribution of user effort, both across scenarios and separately, in Figure 5(B). Here we see relatively low scores for effort overall, consistent with high levels of satisfaction.[4] When we break down the effort over the scenarios, a similar picture emerges as with user satisfaction: participants spend more effort on search tasks, especially structured search.

We now perform a correlation analysis of user satisfaction and its components. Table 2 presents the correlation of user satisfac-

[4]To be precise, this is based on the response to the question 'was a lot of effort was required to complete the task?', measured on a Likert scale, where low scores indicate disagreement with the statement, hence that not much effort was required.

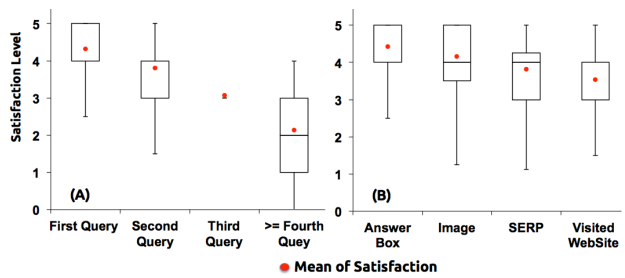

Figure 6: User satisfaction in the web search scenario: satisfaction over the number of queries that users run to find a required answer (A), and over where users find a required answer (B). The mean is represented by the dot and the median is the horizontal line.

tion with **(1)** speech recognition quality (ASR), **(2)** task completion (participants indicate if they are able to complete the suggested task), and **(3)** effort spent (participants report the perceived effort to complete the task). We also look at the correlation between effort and completion. An obvious finding is that user satisfaction depends on ASR quality which is consistent with previous research [25]. Hence ASR quality is a key component of user satisfaction. We find a more interesting pattern for task completion: there is a high correlation with satisfaction for device control, but a low correlation for the information seeking scenarios. This suggests that users are able to find the required information and complete their tasks even in cases where their user satisfaction is suboptimal. And the strong negative correlation between satisfaction and effort shows that users spend a considerable amount of effort to complete their task.

This has important methodological consequences: we cannot equate 'success' in terms of task completion with user satisfaction for the informational scenarios, and have to incorporate the effort taken as a key component of user satisfaction across the different intelligent assistant scenarios. This finding is in line with recent work on task complexity or difficulty and effort, which postulates that satisfaction is low (high) for tasks that take more (less) effort than expected [29]. In addition, ASR quality is of obvious influence on user satisfaction. However, speech recognition is improving constantly and reached the levels that users can recover from misrecognition within a dialogue and still complete their task, at the cost of some extra effort and frustration.

5.2 Good Abandonment for Web Search

We continue with investigating our **RQ 4:** *How to characterize 'abandonment' in the web search scenario?* Whilst intelligent assistants can encourage highly interactive sessions, many results are provided as answers in speech or on the screen, requiring no further interaction of the user (e.g. no need to open a web page and read further to extract the requested information). Hence many sessions stop without an explicit user action, making it hard to discern good and bad search abandonment from interaction log data.

We analyze the phenomenon of *'good abandonment'* from two perspectives: **(1)** the session length and **(2)** where users find the answer addressing their intent. Figure 6(A) presents the dependency of user satisfaction and how much effort was required to find an answer. Effort is associated with the number of queries that participates issued to find the required information. Our observations suggest that user satisfaction is higher if users use fewer queries to reach their goal. Figure 6(A) suggests that if users cannot find an answer after their first query their satisfaction goes down dramatically. Longer sessions lead to user frustration; however, task completion levels are high for the web search scenario, indicating that unnecessary effort was spent in completing the task.

Figure 7: A distribution of overall user satisfaction for different types of tasks: 'simple' tasks, and 'mission' tasks with two and three objectives.

Figure 6(B) shows the dependency of user satisfaction on the place where users find the desired answer. Furthermore, users are more satisfied if they can find a required result directly ('Answer Box' and 'Image') without the need to interact with the SERP such as **(1)** finding an answer in snippets ('SERP'); **(2)** clicking on SERP ('Visited Website'). Hence, cases without further interaction ('Answer Box' and 'Image') lead to higher levels of satisfaction than those requiring interaction ('SERP' and 'Visited Website'). This has important methodological consequences: we have to consider cases of *'good abandonment'*. To measure user satisfaction in this case we need to investigate the other forms of interaction signals that are not based on clicks, such as touch or swipe interactions.

5.3 Analyzing Structured Search Dialogues

We now investigate our **RQ 5:** *How does query-level satisfaction relate to overall user satisfaction for the structured search dialogue scenario?* Structured search dialogues are complex interactions with a longer session and different sub-tasks and changes of focus within the same context. This is very different from traditional search in the query-response paradigm, and session context becomes of crucial importance.

We start our analysis of the collected user interactions with structured search dialogues by introducing the satisfaction distribution for the different types of tasks presented in Figure 7. We see that users are more satisfied with the simple tasks (A), where almost all participants give the highest possible rating. The 'mission' tasks (B and C), that are more complex have a less skewed satisfaction distribution. This immediately shows the complexity of context in structured search dialogues: when viewed independently the quality of the results is comparable for each step of the interaction, and the high levels of satisfaction for the simple task confirm that the quality is high, yet the satisfaction levels go down considerably when tasks are of increasing complexity. This suggests that the intelligent assistant loses context of a conversation, and requires more effort and interaction to restart the dialogue and get back on track. This observation is in line with our previous finding that the amount of effort users spend on a task is a principal component of user satisfaction.

We look now in greater detail at the mission tasks that contain 2 or more sub-tasks, and try to find out how overall user satisfaction is related to user satisfaction per sub-task. Table 3 presents the correlation between the overall *task*-level satisfaction and the minimum, mean, and maximum *query*-level satisfaction per sub-task. The results suggest that overall user satisfaction with the 'mission' tasks depends more on either user frustration—some sub-task results in low satisfaction and frustration dragging down the overall satisfaction fast—or on user success—high levels of satisfaction with the

Table 3: Correlations of overall task user satisfaction and different summations over sub-tasks satisfaction. All presented results are statistical significant ($p < 0.05$)

Measures	Mission tasks
Overall SAT vs. *Average* Sub-task SAT	0.50
Overall SAT vs. *Minimum* Sub-task SAT	0.69
Overall SAT vs. *Maximum* Sub-task SAT	0.71

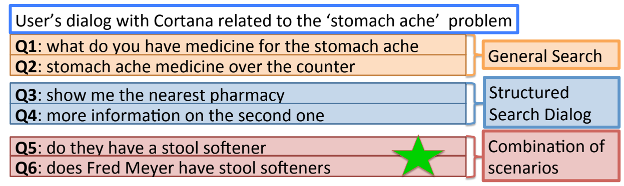

Figure 8: Example of a mixed dialogue.

main sub-task solving the problem lead to high levels of overall satisfaction. This has important methodological consequences: user satisfaction with the structured search dialogues cannot be measured by averaging over satisfaction with sub-tasks, suggesting that task-level satisfaction is different from sub-task or query-level satisfaction, and session-level features are a crucial component.

To summarize, this section show the main results of the user study. We first looked at user satisfaction and found high levels of satisfaction throughout, but important differences between the scenarios on the factors contributing to overall satisfaction: the device control scenario completion correlates well with user satisfaction—it either worked or it did not—but the informational scenarios effort has a much higher correlation with user satisfaction than completion. We then looked in detail at the web search scenario. We found satisfaction dropping fast with the number of issued queries. We also found that direct answers (not requiring interaction) had higher levels of user satisfaction than SERP or web-page results (requiring further interaction) making 'good abandonment' a frequent case and necessitating to take other features (e.g., touch, swipe, acoustic) into account to discern good and bad abandonment. Finally, we zoomed in on the structured search dialogues, and found high level of satisfaction per sub-task but a drop in overall satisfaction for 'mission' tasks with multiple sub-tasks addressing different aspects, showing the importance of preserving session context and demonstrating that task-level satisfaction cannot be reduced to query- or impression-level satisfaction.

6. DISCUSSION AND CONCLUSIONS

This paper aimed to answer the following main research question: *What determines user satisfaction with intelligent assistants?*, by investigating key aspects that determine user satisfaction for different scenarios of intelligent assistant usage. Our first research question was: **RQ 1:** *What are characteristic types of scenarios of use?* We proposed three main types of scenarios of use: **(1)** device control; **(2)** web search; and **(3)** structured search dialogue. The scenarios were identified on the basis of three factors: their proportional existence in the logs of a commercial intelligent assistant; the way requests are handled at the intelligent assistant backend (e.g. user requests are redirected to the different services and they serve different interfaces); and the way scenarios were defined in previous works [25]. Next, we investigated: **RQ 2:** *How can we measure different aspects of user satisfaction?* We designed a series of user studies tailored to the three scenarios of use, with questionnaires on variables potentially related to user satisfaction. The used tasks were based on an extensive analysis of logs of a commercial intelligent assistant.

The data collected in the user study was used to investigate the remaining research questions. First, we looked at: **RQ 3:** *What are key factors determining user satisfaction for the different scenarios?* We collected participant's responses on their satisfaction with the task, their ability to complete a task, and the estimated effort it took. Our main conclusion is that effort is a key component of user satisfaction across the different intelligent assistants scenarios. Second, we focused on the web search interactions: **RQ 4:** *How to characterize 'abandonment' in the web search scenario?* We clearly demonstrated a 'presence' of 'good abandonment' in the web search scenario, and concluded that to measure user satisfaction we need to investigate the other forms of interaction signals that are not based on clicks or reformulation. Third, we zoomed in on the structured dialogue interactions: **RQ 5:** *How does query-level satisfaction relate to overall user satisfaction for the structured search dialogue scenario?* We looked at user satisfaction as 'a user journey towards an information goal where each step is important,' and showed the importance of session context on user satisfaction. Our experimental results show that user satisfaction cannot be measured by averaging over satisfaction with sub-tasks. Hence, frustration with some steps in a user's 'journey' can greatly affect their overall satisfaction.

Our general conclusion is that the factors contributing to overall satisfaction with a task are different between the scenarios. Task completion is highly correlated with user satisfaction for the device control scenario—it either worked or it did not. For information seeking scenarios, user satisfaction is more related to effort than task completion. We demonstrated that task-level satisfaction cannot be reduced to query or impression-level satisfaction for information seeking scenarios.

Research on intelligent assistants for mobile devices is a new area, and this paper addresses some of the important first steps. This work can be extended in two main directions. First, our taxonomy of three types of scenarios could be extended in various ways. In the logs we noticed that users use a mix of scenarios in order to satisfy their information needs. Consider for example the dialogue in Figure 8, in which the user combined multiple different scenarios in order to accomplish his/her task: The user started by using general web search (Step 1: $Q_1 \rightarrow Q_2$) to get information about his/her problem. Then he/she used the structured search dialogue (Step 2: $Q_3 \rightarrow Q_4$) to find a pharmacy. Afterwards, he/she attempted to combine the information from the prior steps through complex requests (Step 3: $Q_5 \rightarrow Q_6$). Unfortunately, this led to dissatisfaction as the intelligent assistant failed to process Step 3. Therefore, it is essential to study user satisfaction when users use the mix of scenarios. Second, we found that typical behavioral signals in interaction logs (e.g., clicks) are not sufficient to infer user satisfaction with intelligent assistants. Going forward, therefore, it will be important to make use of other types of interactions such as touch or swipe, or acoustic signals to predict user satisfaction. It has been shown [13, 25, 33] that these signals are promising to detect user satisfaction with intelligent assistants and hold the potential to construct accurate predictions of task-level user-satisfaction based on behavioral data. Ultimately, such signals can be used in production systems to improve the quality of human interaction with intelligent assistants.

Acknowledgments

We thank Sarvesh Nagpal and Toby Walker for the help in collecting the internal API data for the user study. We also thank the participants in the study.

References

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *SIGIR*, 2011.

[2] M. Ageev, D. Lagun, and E. Agichtein. Improving search result summaries by using searcher behavior data. In *SIGIR*, pages 13–22, 2013.

[3] E. Agichtein, E. Brill, and S. T. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, 2006.

[4] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *SIGIR*, pages 773–774, 2007.

[5] P. Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf. Res. (IRES)*, 8(3), 2003.

[6] A. Chuklin and P. Serdyukov. How query extensions reflect search result abandonments. In *SIGIR*, pages 1087–1088, 2012.

[7] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *WWW (Companion Volume)*, pages 483–484, 2012.

[8] A. Diriye, R. White, G. Buscher, and S. T. Dumais. Leaving so soon?: understanding and predicting web search abandonment rationales. In *CIKM*, pages 1025–1034, 2012.

[9] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, and C. L. F. Diaz. Towards recency ranking in web search. In *WSDM*, pages 11–20, 2010.

[10] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR*, pages 34–41, 2010.

[11] S. Fox, K. Karnawat, M. Mydland, S. T. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst. (TOIS)*, 23(2):147–168, 2005.

[12] Google Inc. Teens use voice search most, even in bathroom, google's mobile voice study finds, 2015. http://prn.to/1sfiQRr.

[13] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Towards estimating web search result relevance from touch interactions on mobile devices. In *CHI Extended Abstracts 2013*, pages 1821–1826, 2013.

[14] A. Hassan. A semi-supervised approach to modeling web search satisfaction. In *SIGIR*, pages 275–284, 2012.

[15] A. Hassan and R. W. White. Personalized models of search satisfaction. In *CIKM*, pages 2009–2018, 2013.

[16] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *WSDM*, pages 221–230, 2010.

[17] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: query reformulation as a predictor of search satisfaction. In *CIKM*, pages 2019–2028, 2013.

[18] A. Hassan Awadallah, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: disambiguating long search sessions. In *WSDM*, pages 53–62, 2014.

[19] L. P. Heck, D. Hakkani-Tür, M. Chinthakunta, G. Tür, R. Iyer, P. Parthasarathy, L. Stifelman, E. Shriberg, and A. Fidler. Multi-modal conversational search and browse. In *SLAM@INTERSPEECH*, pages 96–101, 2013.

[20] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, 2005.

[21] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst. (TOIS)*, 20(4):422–446, 2002.

[22] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *ECIR*, pages 4–15, 2008.

[23] J. Jiang, W. Jeng, and D. He. How do users respond to voice input errors?: lexical and phonetic query reformulation in voice search. In *SIGIR*, pages 143–152, 2013.

[24] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM*, 2015.

[25] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. G. Kulkarni, and O. Z. Khan. Automatic online evaluation of intelligent assistants. In *WWW*, pages 506–516, 2015.

[26] T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

[27] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, 2005.

[28] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval (FTIR)*, 3(1-2):1–224, 2009.

[29] D. Kelly. When effort exceeds expectations: A theory of search task difficulty (keynote). In *SCST'15: Proceedings of the First International Workshop on Supporting Complex Search Tasks*, volume 1338 of *CEUR Workshop Proceedings*, 2015.

[30] Y. Kim, A. Hassan, R. W. White, and Y.-M. Wang. Playing by the rules: mining query associations to predict search performance. In *WSDM*, pages 133–142, 2013.

[31] J. Kiseleva, E. Crestan, R. Brigo, and R. Dittel. Modelling and detecting changes in user satisfaction. In *CIKM*, pages 1449–1458, 2014.

[32] J. Kiseleva, J. Kamps, V. Nikulin, and N. Makarov. Behavioral dynamics from the serp's perspective: What are failed serps and how to fix them? In *CIKM*, 2015.

[33] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR*, pages 113–122, 2014.

[34] J. Li, S. B. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR*, pages 43–50, 2009.

[35] M. F. McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.

[36] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna. Quality estimation for automatic speech recognition. In *COLING*, pages 1813–1823, 2014.

[37] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26:321–343, 1975.

[38] T. Saracevic, P. B. Kantor, A. Y. Chamis, and D. Trivison. A study of information seeking and retrieving. I. background and methodology. II. users, questions and effectiveness. III. searchers, searches, overlap. *Journal of the American Society for Information Science and Technology*, 39:161–176; 177–196; 197–216, 1988.

[39] M. Shokouhi and Q. Guo. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *SIGIR*, pages 695–704, 2015.

[40] G. Tür. Extending boosting for large scale spoken language understanding. *Machine Learning (ML)*, 69(1):55–74, 2007.

[41] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Techware: Spoken language understanding resources [best of the web]. *IEEE Signal Process. Mag. (SPM)*, 30(3):187–189, 2013.

[42] G. Tür, Y.-Y. Wang, and D. Z. Hakkani-Tür. Understanding spoken language. *Computing Handbook*, 3rd ed(41):1–17, 2014.

[43] P. Vakkari. Task-based information searching. *ARIST*, 37:413–464, 2003.

[44] W. Wahlster. Smartkom: Foundations of multimodal dialogue systems. *Springer*, 2006.

[45] Y. Wang and E. Agichtein. Query ambiguity revisited: Clickthrough measures for distinguishing informational and ambiguous queries. In *HLT-NAACL*, pages 361–364, 2010.

[46] B. Wildemuth, L. Freund, and E. G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation*, 70:1118–1140, 2014.

[47] T. Wilson. Models in information behaviour research. *Journal of Documentation*, 55(3):249 – 270, 1999.

[48] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In *CIKM*, pages 91–100, 2014.